#4

Office de la propriété intellectuelle du Canada

Un organisme d'Industrie Canada

Canadian Intellectual Property Office

An Agency of Industry Canada

# Bureau canadien des brevets

*Certification*

# Canadian Patent Office

*Certification*

La présente atteste que les documents ci-joints, dont la liste figure ci-dessous, sont des copies authentiques des documents déposés au Bureau des brevets.

This is to certify that the documents attached hereto and identified below are true copies of the documents on file in the Patent Office.

Specification and Drawings, as originally filed, with Application for Patent Serial No: **2,331,474**, on January 19, 2001, by **STERGIOS V. ANASTASIADIS**, for "Stride-Based Disk Space Allocation Scheme".

S Gregoire

Agent certificateur/Certifying Officer

February 18, 2002

Date

Canada

OPIC CIPO

Office de la propriété intellectuelle du Canada

Un organisme d'Industrie Canada

Canadian Intellectual Property Office

An Agency of Industry Canada

#4

# *Bureau canadien des brevets*

*Certification*

# *Canadian Patent Office*

*Certification*

La présente atteste que les documents ci-joints, dont la liste figure ci-dessous, sont des copies authentiques des documents déposés au Bureau des brevets.

This is to certify that the documents attached hereto and identified below are true copies of the documents on file in the Patent Office.

Specification and Drawings, as originally filed, with Application for Patent Serial No: **2,331,474**, on January 19, 2001, by **STERGIOS V. ANASTASIADIS**, for "Stride-Based Disk Space Allocation Scheme".

S. Gregoire

Agent certificateur/Certifying Officer

February 18, 2002

Date

Canada

OPIC CIPO

# Stride-Based Disk Space Allocation Scheme

## Field of Invention

This invention relates to disk space allocation for mass storage servers and in particular to a stride-based disk space allocation scheme. In addition, it relates to methods of striping data across multiple disks for continuous media streaming.

## Background of the Invention

With the installed network bandwidth tripling every year, scalable network servers are becoming the dominating bottleneck in the wide deployment of broadband services over the Internet. A fundamental problem is the potential scalability in the context of network servers of variable bit rate (video) streams.

Spatial and temporal compression have made practical the storage and transfer of digital video streams with acceptable quality. Standardization through the MPEG-2 specification has facilitated widespread distribution and use of compressed video content in a range of applications from studio post-production editing to home entertainment (e.g. Digital Versatile Disks). Although MPEG-2 streams can optionally be encoded in constant bit rate, it has been shown that equally acceptable quality can be achieved using variable bit rate encoding with average bit rates reduced by 40%.

Vast storage and bandwidth capacity requirements of even compressed video streams make it necessary to stripe video files across multiple disks. Assuming that a media storage server serves requests for several different stream files, appropriate striping makes it possible to scale the number of supported streams to the limit of the server resources, independently of the particular stream files being requested by the clients. This becomes possible by retrieving different parts of each stream file from different disks, thus restricting the degree of imbalance in utilization among the disks.

Several media server designs either i) support only constant bit rate streams, ii) make resource reservations assuming a fixed bit rate for each stream, or iii) have only been demonstrated to work with constant bit rate streams. It has been found that both load imbalance across disks and disk overhead caused stream striping to be efficient only on disk arrays of limited size. The scalability of network servers to provide video streams is a fundamental problem.

It is therefore an aspect of an object of the present invention for providing a new disk space allocation technique and striping policies that increase system throughput and improve scalability.

## Brief Description of the Drawings

In the accompanying drawings: Figures 1 to 16

## Detailed Description of the Preferred Embodiments

Referring to the drawings and initially to Figure 1, there is illustrated a distributed media server in accordance with an embodiment of the present invention comprises disks for storing stream data, transfer nodes, admission control nodes, and a schedule database containing scheduling information. The streams are compressed according to the MPEG-2 specification, with constant quality quantization parameters and variable bit rates. Clients with appropriate stream decoding capability send playback requests and receive stream data via a high-speed network, as shown in Figure 1. Alternately, the compression can be of any type that supports variable bit rates.

In the media server, stream data are retrieved from the disks and sent to the clients through the Transfer Nodes. Both the admission control and the data transfers make use of stream scheduling information maintained in the Schedule Database.

The media server is operated using the server-push model, but other models are possible. When a playback session starts, the server periodically sends data to the client until either the end of the stream is reached, or the client explicitly requests suspension of the playback. The server-push model reduces the control traffic from the client to the server and facilitates resource reservation at the server side, when compared to a client-pull model. The data transfers occur in rounds of fixed duration $T_{round}$: in each round, an appropriate amount of data is retrieved from the disks into a set of server buffers reserved for each active client. Concurrently, data are sent from the server buffers to the client through the network interfaces. Round-based operation is used in media servers in order to keep the reservation of the resources and the scheduling-related bookkeeping of the data transfers manageable.

Due to the large amount of network bandwidth required for this kind of service, preferably the server is connected to a high-speed network through multiple network interfaces. The amount of stream data periodically sent to the client is determined by the decoding frame rate of the stream and the resource management policy of the network. A reasonable policy is to send to the client during each round the amount of data that will be needed for the decoding process of the next round; any other policy that does not violate the timing requirements and buffering constraints of the decoding client is also acceptable.

The stream data are stored across multiple disks, as shown in Figure 1. Every disk is connected to a particular Transfer Node, through the Storage Interconnect, which is either i) a standard I/O channel (e.g. Small Computer System Interface), ii) standard network storage equipment (e.g. Fibre-Channel), or iii) a general purpose network (as with Network-Attached Secure Disks). Alternately, the file server functionality can be offloaded to network-attached disks.

The Transfer Nodes are computers responsible for scheduling and initiating all data accesses from the attached disks. Data arriving from the disks are temporarily staged in the Server Buffer memory of the Transfer Node before being sent to the client through the high-speed network. The bandwidth of the system bus (such as the Peripheral Component Interconnect) is the critical resource within each Transfer Node that essentially defines the number and the capacity of the attached network or I/O channel interfaces.

Playback requests arriving from the clients are initially directed to an Admission Control Node, where it is decided whether enough resources exist to activate the requested playback session either immediately or within a few rounds. If a new playback request is accepted, commands are sent to the Transfer Nodes to begin the appropriate data accesses and transfers. The computational complexity of the general stream scheduling problem is combinatorial in the number of streams considered for activation and the number of reserved resources. As the acceptable initiation latency is limited, a simple scheduling approach with complexity linear with the number of rounds of each stream and the number of reserved resources, is used. The admission control is distributed across multiple processors as shown in Figure 1, and concurrency control issues that potentially arise are also taken into account.

The Schedule Database maintains information on the amount of data that needs to be retrieved during each round for each stream and on which disks this data is stored. It also specifies the amount of buffer space required and the amount of data sent to the client by the Transfer Nodes during each round. The scheduling information is generated before the media stream is first stored and is used for both admission control and for controlling data transfers during playback. Since this information changes infrequently, it is replicated to avoid potential bottlenecks.

Referring to Figure 2, there is illustrated a stride-based allocation of disk space shown on one disk. A stream is stored in a sequence of generally non-consecutive fixed-size strides with a stride possibly containing data of more than one round. Sequential requests of one round are smaller than the stride size and thus require at most two partial stride accesses.

In stride-based allocation, disk space is allocated in large, fixed-sized chunks called strides, which are chosen larger than the maximum stream request size per disk during a round. The stored streams are accessed sequentially according to a predefined (albeit variable) rate; therefore, the maximum amount of data accessed from a disk during a round for a stream is known a priori. Stride-based allocation eliminates external fragmentation, while internal fragmentation remains negligible because of the large size of the streams, and because a stride may contain data of more than one round as shown in Figure 2.

When a stream is retrieved, only the requested amount of data is fetched to memory, and not the entire stride (that is sequentially allocated on the disk platters). Another advantage of stride-based allocation is that it sets an upper-bound on the estimated disk access overhead during retrieval. Since the size of a stream request never exceeds the stride size during a round, at most two partial stride accesses is required to serve the request of a round on each disk. This avoids the arbitrary number of actuator movements required by prior allocation methods.

While storing the data of each disk request contiguously would reduce the disk overhead to a single seek and rotation delay (instead of two at most), the overhead for storage management (bookkeeping) of large highly utilized disks could become significant. An advantage of the present invention is the reduction of overhead for storage management.

In the sequence definitions that follow, a zero value is assumed outside the specified range.

In a system with D functionally equivalent disks, the stream Network Sequence, Sn, of length Ln defines the amount of data, Sn[i], 1=< i =< Ln, that the server sends to a particular client during round i after its playback starts. Similarly, the Buffer Sequence Sb of length Lb = Ln + 1 defines the server buffer space, Sb(i), =< i =< Lb, occupied by the stream data during round i. The Disk Sequence Sd of length Ld = Ln defines the total amount of data Sd(i), =< i =< Ld - 1, retrieved from all the disks in round i for the client.

The data are stored on the disks in strides. The stride sizc Bs is a multiple of the logical block size Bl, which is a multiple of the physical sector size Bp of the disk. Both disk transfer requests and memory buffer reservations are specified in multiples of the logical block size Bl. After taking into account logical block quantization issues, the disk sequence Sd can be derived from the network sequence Sn as follows: If

$$K^d(i) = \left\lceil \frac{\sum_{0 \leq j \leq i} S_n(j+1)}{B_l} \right\rceil$$

specifies the cumulative number of blocks Bl retrieved through round i, then

$$S_d(i) = (K^d(i) - K^d(i-1)) \cdot B_l.$$

The Disk Striping Sequence Smd of length Ld determines the amount of data Smd(i,k), 0 =< i =< Ld - 1, that are retrieved from the disk k, 0 =< k =< D-1, in round i. It is generated from the Disk Sequence Sd, according to the striping policy used.

Each disk has edge to edge seek time TfullSeek, single track seek time TtrackSeek, average rotation latency TavgRot, and minimum internal transmission rate Rdisk. The stride-based disk space allocation policy enforces an upper bound of at most two disk arm movements per disk for each client per round. The total seek distance is also limited using a CSCAN disk scheduling policy. Let Mi be the number of active streams during round i of the system operation. Where the playback of stream j, 1 =< j =< Mi, is initiated at round lj of system operation, then, the total access time on disk k in round i of the system operation has an upper-bound of:

$$T_{disk}(i,k) = 2T_{fullSeek} + 2M_i \cdot (T_{trackSeek} + T_{avgRot}) + \sum_{j=1}^{M_i} S^j_{md}(i - l_j, k)/R_{disk}$$

where Smd^j is the disk striping sequence of client j. TfullSeek is counted twice due to the disk arm movement from the CSCAN policy, while the factor two of the second term is due to the stride-based method. The first term should be accounted for only once in the disk time reservation structure of each disk. Then, each client j incurs an additional maximum access time of

$$T^{j}_{disk}(i,k) = 2 \cdot (T_{trackSeek} + T_{avgRot}) + S^{j}_{md}(i - l_j, k)/R_{disk}$$

on disk k during round i, when Smd^j(i-lj,k) > 0, and zero otherwise.

If Rnet is the total high-speed network bandwidth available to the server, then the corresponding network transmission time reserved for client j in round i becomes T^jnet(i) = Sn^j(i-lj) / Rnet, where Sn is the Network Sequence of client j. The total server memory buffer reserved for client j in round i becomes B^j(i) = Sb^j(i-lj), where Sb is the Buffer Sequence of client j. Although, the above expressions for T^jnet(i) and B^j(i) are sufficient for the needs of the present embodiments, accounting for available network bandwidth and buffer memory within each individual Transfer Node may require them to be split into appropriate sub-expressions.

The reservations of transfer time on each network interface and buffer space on each transfer node are more straightforward, and are based on the Network Striping Sequence and Buffer Striping Sequence, respectively.

In traditional storage systems, data access patterns are relatively hard to predict, making it difficult to determine optimal disk striping parameters. However, with read-only sequential access being the common case for video streaming, it is possible to predict to some degree the expected system load requirements during retrieval, making it possible to determine appropriate disk striping parameters a priori for the storage and retrieval of the data. The present invention includes exploiting this characteristic of stored video streams.

Referring to Figure 3, there is illustrated the data requirements of twenty consecutive rounds (one second each), in an MPEG-2 clip. Referring to Figure 4, there is illustrated data accesses for the clip of Figure 3 using alternative striping techniques over two disks. With Fixed-Grain Striping, the needed blocks of size Bf are retrieved round-robin from the two disks every round. In Variable-Grain Striping, a different disk is accessed in each round, according to the byte requirements of the original clip. In Group-Grain Striping with G=2, stream data worth of two rounds are accessed from a different disk every two rounds.

With Fixed-Grain Striping, data are striped round-robin across the disks in blocks of a fixed size Bf, a multiple of the logical block size Bl defined previously. During each round, the required number of blocks are accessed from each disk. An example of Fixed-Grain Striping is shown in Figure 4(a). In the definition below, modD denotes the remainder of the division with D, and divD denotes the integer quotient of the division with D. The equation

$$K^f(i) = \left\lceil \frac{\sum_{0\le j\le i} S_d(j)}{B_f} \right\rceil,$$

specifies the cumulative number of blocks Bf retrieved through round i for a specific client. When

$$K^f_{divD}(i) -$$

$$K^f_{divD}(i-1) = 0,$$

all blocks accessed for the client during round i lie on the same stripe of blocks. Then, the striping sequence Smd^f is equal to:

$$S^f_{md}(i,k) = D^f_0(i,k) \cdot B_f$$

where

$$D^f_0(i,k) = \begin{cases} 1, & \text{if } K^f_{modD}(i-1) < k_{modD} \le K^f_{modD}(i) \\ 0, & \text{otherwise,} \end{cases}$$

specifies the particular disks that need to be accessed at most once for the stream. When

$$K^f_{divD}(i) - K^f_{divD}(i-1) > 0,$$

the blocks accessed for the client during round i lie on more than one stripe, and the striping sequence becomes

$$S^f_{md}(i,k) = (K^f_{divD}(i) - K^f_{divD}(i-1) - 1) \cdot B_f + D^f_{>0}(i,k) \cdot B_f,$$

where

$$D^f_{>0}(i,k) = \begin{cases} 2, & \text{if } K^f_{modD}(i-1) < k_{modD} \leq K^f_{modD}(i) \\ 1, & \text{if } k_{modD} > max(K^f_{modD}(i-1), K^f_{modD}(i)\ )' \\ 1, & \text{if } k_{modD} \leq min(K^f_{modD}(i-1),\ K^f_{modD}(i)) \\ 0, & \text{otherwise.} \end{cases}$$

The first term in the Equation accounts for blocks in stripes fully accessed (i.e., all D blocks, where D is the number of disks), while the second term accounts for blocks of stripes partially accessed in round i (i.e., fewer than D blocks).

With Variable-Grain Striping, the data retrieved during a round for a client are always accessed from a single disk round-robin, as shown in Figure 4(b). The corresponding striping sequence becomes:

$$S^v_{md}(i,k) = (K^v(i) - K^v(i-1)) \cdot B_l,$$

when i mod D=k, with

$$K^v(i) = \left\lceil \frac{\sum_{0 \leq j \leq i} S_d(j)}{B_l} \right\rceil,$$

and Smd^v(i,k)=0 when i mod D not equal k. Therefore, the Disk Sequence determines the particular single disk accessed and the exact amount of data retrieved during each round.

Variable-Grain Striping is a special case (with G=1) of a method herein called Group-Grain Striping, where the amount of data required by a client over G rounds is retrieved every Gth round from one disk that changes round robin (see Figure 4(c), noting that the y-axis uses a different scale). The parameter G, G >= 1, is called Group Size. The striping sequence for Group-Grain Striping is equal to:

$$S^{g}_{md}(i,k) = (K^{v}(i+G-1) - K^{v}(i-1)) \cdot B_l$$

when i mod G=0 AND (i div G) mod D=k, and Smd^g(i,k) = 0 otherwise.

As G increases, the fewer disk accesses lead to reduced disk overhead (although the access time per request is increased). On the other hand, the fixed round spacing between subsequent requests for a stream, basically divides the server into G virtual servers. The fixed group size G guarantees that two streams started from the same disk at rounds i and j with i not equal j (mod G), do not have any disk transfers in a common round. This is different than increasing Bf in Fixed-Grain Striping, where accesses from different streams can randomly coincide on the same disk in the same round, resulting in the system saturating with fewer streams. Increasing G for a particular round time is advantageous with future expected changes in disk technology.

Alternately, aggregation of disk transfers can also be achieved with an appropriate increase of round time. However, this could directly affect the responsiveness of the system by potentially increasing the initiation latency of each playback. Longer round time would also increase the required buffer space.

The above disclosure generally describes the present invention. A more complete understanding can be obtained by reference to the following specific Examples. These Examples are described solely for purposes of illustration and are not intended to limit the scope of the invention. Changes in form and substitution of equivalents are contemplated as circumstances may suggest or render expedient. Although specific terms have been employed herein, such terms are intended in a descriptive sense and not for purposes of limitation.

## Examples

The examples are described for the purposes of illustration and are not intended to limit the scope of the invention.

## A Media Server System

A media server system was built in order to evaluate the resource requirements of the different striping techniques. The modules were implemented in about 12,000 lines of C++/ Pthreads code on AIX4.1, and ran on a single node. The code was linked either to the University of Michigan DiskSim disk simulation package, which incorporated advanced features of modern disks such as on-disk cache and zones for simulated disk access time measurements, or to hardware disks through their raw device interfaces. The indexing metadata were stored as regular Unix files, and during operation were kept in main memory. The MPEG-2 decoder from the MPEG Software Simulation Group was used for stream frame size identification.

The basic responsibilities of the media server included file naming, resource reservation, admission control, logical to physical metadata mapping, buffer management, and disk and network transfer scheduling.

With appropriate configuration parameters, the system operated at different levels of detail. In Admission Control mode, the system receives playback requests, does admission control and resource reservation but no actual data transfers take place. In Simulated Disk mode, all the modules are functional, and disk request processing takes place using the specified DiskSim disk array.

The system was primarily used in Admission Control mode (except for our validation study, where the system was used in Simulated Disk mode). The Admission Control module used circular vectors of sufficient length to represent the allocated time of each disk, the network time, and the buffer space respectively. On system startup, the disk time vectors are initialized to 2•TfullSeek, while the network time and buffer space are initially set to zero. When a new stream request arrived, admission control is performed by checking against current available resources. In particular, the total service time of each disk in any round may not exceed the round duration, and the total network service time may also not exceed the round duration, while the total occupied buffer space may be no longer than the server buffer capacity. If the admission control test was passed, the resource sequences of the stream are added to the corresponding vectors of the module, and the stream is scheduled for playback.

Performance Evaluation Method

The playback initiation requests arrived independently of one another, according to a Poisson process. The system load was controlled through the mean arrival rate λ of playback initiation requests. Assuming that disk transfers form the bottleneck resource, in a perfectly efficient system there is no disk overhead involved in accessing disk data. Then, the maximum arrival rate λ was chosen to equal λmax of playback initiation requests, that corresponds to system load 100%, to be equal to the mean service rate with which stream playbacks would complete in that perfectly efficient system. This makes it possible to show the performance benefit of arbitrarily efficient data striping policies. Subsequently, the mean service rate μ, expressed in streams per round, for streams of data size Stot bytes becomes: μ = D•Rdisk•Tround/Stot. Correspondingly, the system load ρ, was set equal to: ρ= λ/μ =< 1, where λ =< λmax = μ.

Another important decision had to do with the admission control process. When a playback request arrived, it is checked to determine if available resources existed for every round during playback. The test considered the exact data transfers of the requested playback for every round and also the corresponding available disk transfer time, network transfer time and buffer space in the system. If the next round failed this test, it is repeated until the first future round is found, where the requested playback can be started with guaranteed sufficiency of resources.

The lookahead distance Hl was defined as the number of future rounds that are considered as candidate rounds for initiating the stream for each request before it is rejected. Playback requests not accepted were turned away rather than being kept in a queue. Practically, a large lookahead distance allowed a long potential waiting time for the initiation of the playback. It cannot be unlimited in order for the service to be acceptable by the users. On the other hand, setting the lookahead distance too small can prevent the system from reaching full capacity.

The basic lookahead distance Hl^basic was set to be equal to the mean number of rounds between request arrivals Hl^basic= 1/λ. Setting Hl = Hl^basic allows the system to consider for admission control the number of upcoming rounds that will take (on average) for another request to arrive. More generally, a lookahead factor Fl is defined as the fraction Fl = Hl/Hl^basic.

As the basic performance metric, the expected number of active playback sessions that can be supported by the server was chosen. The objective was to make this number as high as possible.

| Content Type | Avg Bytes per rnd | Max Bytes per rnd | CoV per rnd |
|---|---|---|---|
| Science Fiction | 624935 | 1201221 | 0.383 |
| Music Clip | 624728 | 1201221 | 0.366 |
| Action | 624194 | 1201221 | 0.245 |
| Talk Show | 624729 | 1201221 | 0.234 |
| Adventure | 624658 | 1201221 | 0.201 |
| Documentary | 625062 | 625786 | 0.028 |

**Table 1.** We used six MPEG-2 video streams of 30 minutes duration each. The coefficient of variation shown in the last column changes according to the content type.

| Seagate Cheetah ST-34501 | |
|---|---|
| Data Bytes per Drive | 4.55 GByte |
| Average Sectors per Track | 170 |
| Data Cylinders | 6,526 |
| Data Surfaces | 8 |
| Zones | 7 |
| Buffer Size | 0.5MByte |
| Track to Track Seek(read/write) | 0.98/1.24 msec |
| Maximum Seek(read/write) | 18.2/19.2 msec |
| Average Rotational Latency | 2.99 msec |
| Internal Transfer Rate<br>Inner Zone to Outer Zone Burst<br>Inner Zone to Outer Zone Sustained | <br>122 to 177 Mbits<br>11.3 to 16.8 MByte |

**Table 2.** Features of the SCSI disk assumed in our experiments.

## Setup

Six different VBR MPEG-2 streams of 30 minutes duration each were used. Each stream had 54,000 frames with a resolution of 720x480 and 24 bit color depth, 30 frames per second frequency, and a $IB^2PB^2PB^2PB^2PB^2$ 15 frame Group of Pictures structure. The encoding hardware that was used allowed the generated bit rate to take values between 1Mbps and 9.6Mbps. Although the Main Profile Main Level MPEG-2 specification allows bitrates up to 15Mbit/sec, there is a typical point of diminishing returns (no visual difference between original and compressed video) at 9Mbit/sec. The DVD specification sets a maximum allowed MPEG-2 bitrate of 9.8Mbit/sec. Statistical characteristics of the clips are given in Table 1, where the coefficients of variation lie between 0.028 and 0.383, depending on the content type. In the mixed basic benchmark, the six different streams were submitted in a round-robin fashion. Where necessary, the results from individual stream types are also shown.

For the measurements, Seagate Cheetah ST-34501 SCSI disks were assumed, with the features shown in Table 2. Except for the storage capacity, which can reach 73GB in the latest models, the rest of the performance numbers are typical of today's high-end drives. The logical block size Bl was set to 16,384 bytes, while the physical sector size Bp was equal to 512 bytes. The stride size Bs in the disk space allocation was set to 1,572,864 bytes. The server memory was organized in buffers of fixed size Bl=16,384 bytes each, with a total space of 64MB for every extra disk. The available network bandwidth was assumed to be infinite.

The round time was set equal to one second. A warmup period of 3,000 rounds was used and calculated the average number of active streams from round 3,000 to round 9,000. The measurements were repeated until the half-length of the 95% confidence interval was within 5% of the estimated mean value of the active streams.

Example on Fixed-Grain Striping

In respect of Fixed-Grain Striping, an important feature of this method is the ability to control the disk access efficiency through the choice of block size Bf. As the block size is increased, a larger part of each access is devoted to data transfer rather than mechanical movement overhead. When a stream requests more than one block from a particular disk during a round, a maximum of two contiguous accesses is sufficient with the stride-based disk space allocation used.

As shown in Figure 5, the number of active streams with sixteen disks and the mixed workload increases linearly as the load, ρ, increases from 10% to 50%. At loads higher than 50%, the number of streams that can be supported no longer increases. The additional load beyond 50% translates into a corresponding increase in the number of rejected streams (Figure 6). Since increasing the lookahead factor from 1 to 30 improves the number of streams that can be supported only marginally, for the rest of the experiments, the lookahead factor Fl was set to 1. This corresponds to a lookahead distance of less than 10 rounds, for a system of sixteen disks operating at load ρ = 80%, and half-hour clips of 1GByte each.

For load values ρ= 40\% and ρ= 80%, the number of active streams were measured as the block size increased from Bf = 32,768 to Bf = 1,048,576 bytes at steps of 32,768. As can be seen from Figure 7, at load 80% the number of streams initially increases until

Bf becomes equal to 327,680 and then drops. A similar behavior is noticed at 40%, although the variation in the number of streams is much smaller across different block sizes.

The Admission Control mode that was used for the above experiments allowed the gathering of statistics on system resources reserved for each round during the admission control process. In particular, Figure 8 depicts the maximum and average access time Tdisk(i,k) that was reserved during the measurement period 3,000 =< i < 9,000 for a particular disk (k=0) in a sixteen disk configuration with load ρ= 80%. While the maximum time remains close to 100% across different block sizes, the average time drops from about 90% at Bf=32,768 to less than 50% at Bf=1,048,576.

With the round time set to 1 sec, the average time (normalized by the round time) corresponds to the expected disk utilization and varies depending on the number of disks accessed for a stream every round. Part of it was actuator overhead and decreased as the block size becomes larger. On the other hand, the maximum difference in reserved access times in a round (Avg Diff in Figure 8) increased on average from almost zero to above 60%, with increasing block size Bf. This could be another reason for the decrease in the average reserved time for larger block sizes.

It was also found that the average reserved time (shown in Figure 8 only for Disk 0) remains about the same (typically within 2%) across a disk array. Thus, the access load, on average, was equally distributed across the disks, despite variations from round to round. Furthermore, as the number of disks increases, the average time drops only slightly from 69% with 8 disks to 67% with 16 and 66% with 32 disks (Figure 9). This implied that the capacity of the system increases almost linearly as more disks are added.

The above measurements were repeated varying the number of disks from 4 to 64 (Figure 10). The block size Bf, that maximized the number of streams, was found to remain at Bf = 327,680. At 80% load, the number of streams that could be supported increased from 39.17 with 4 disks to 143.57 with 16 disks and 550.23 with 64 disks. This is within 9-14% of what perfectly linear scalability should achieve. With the load at 40%, the number of streams increased from 30.31 with 4 disks to 504.79 with 64 disk, thus reflecting the improved capacity of the system with increased number of disks at low loads.

With Fixed-Grain Striping, the mixed workload the number of streams is maximized at Bf=327,680 across different number of disks and system load values.

## Example of Variable Grain Striping

In this example, Variable Grain Striping is used. In Figure 11, the performance of Variable-Grain Striping on sixteen disks is shown as the load increases from 10% to 100%. The number of streams grows linearly as the load increases up to 70%. This is significantly higher than the 50% load, where Fixed-Grain Striping flattened out (Figure 5). Loads higher than 70% with Variable Grain Striping only increase the number of rejected streams as shown in Figure 12. As before, a lookahead factor value of Fl = 1 attains more than 95% of the system throughput, and that is the value that is used.

In Figure 13, there is illustrated the reserved disk access time per round. As the number of disks increases, the average reserved time increases from 83% with 8 disks, to 84% with 16 disks, and 85% with 32 disks. The maximum number of sustained streams from 4 to 64 disks were also measured (Figure 10). At a load of 80%, the number of streams increases from 48.11 with 4 disks, to 202.69 with 16 disks and 786.05 with 64 disks. Thus, as the number of disks increases, the number of streams remains within 3% of what perfectly linear scalability should achieve. In addition, the advantage of Variable-Grain Striping over Fixed-Grain Striping increases from 23% with 4 disks to 43% with 64 disks.

In Figure 14, there is illustrated individual stream types in more detail. As the content type changes from Science Fiction to Documentary and the variation in data transfers correspondingly drops, the block size has to be larger in order to maximize the performance of Fixed-Grain Striping. However, the performance remains about the same for the five stream types, and increases only with the Documentary stream. In contrast, Variable-Grain Striping manages to transform even minor decreases in data transfer variation into improved performance. Overall, Variable-Grain Striping maintains an advantage over Fixed-Grain Striping between 11% and 50%.

## Validation is Simulated Disk Mode

In order to keep the computation time reasonable, the previous work were conducted with the system in Admission Control mode, where playback requests arrive leading to corresponding resource reservations, but without actual time measurement of the individual disk transfers. The statistics of the disk time resource reservations is compared with the statistics gathered over the access times of all individual data transfers involved, using the DiskSim representation of the Seagate Cheetah ST-34501 disk. A two-disk array model is used with each disk attached to a separate 20MB/sec SCSI bus, and no contention assumed on the host system bus connecting the two SCSI buses. The statistics are gathered during 6,000 rounds after a warmup period of 3,000 rounds, as before. The mixed workload is used with average number of active streams 21.23 and 23.27 for Fixed-Grain and Variable-Grain Striping, respectively, corresponding to 80% load.

As can be seen from Figure 15, in both the average and maximum case, the reserved disk time is no more than 8% higher than the corresponding measurements using the DiskSim. The difference can be attributed to the fact that the reservation assumes a minimum disk transfer rate and ignores on-disk caching.

Effect of Technology Improvements

To project disk technology improvements for the foreseeable future, the compound growth rates from the past are extended linearly into the future (Table 3). In particular, a 30% increase in internal disk transfer rate per year, and 23% decrease in seek distance is used. The full seek time depends linearly on seek distance, so the decrease is also 23\%. However, a decrease of 12% per year for the track seek time is also assumed, which is dependent on the square root of the seek distance (among other factors more complex to project including settle time). Finally, a rotation speed increase of 12% per year is assumed. The stream types and sizes remaining the same.

| Disk Parameter | Today | 2 Years | 5 Years |
|---|---|---|---|
| Min Transfer Rate (MB/sec) | 11.3 | 19.10 | 41.92 |
| Max Seek Time (msec) | 18.2 | 10.74 | 4.91 |
| Track Seek Time (msec) | 0.98 | 0.76 | 0.51 |
| Avg Rotation Latency (msec) | 2.99 | 2.38 | 1.70 |

**Table 3.** Projection of disk parameter changes in two and five years into the future.

The above compared Fixed-Grain Striping to Variable-Grain Striping, which is a special case of Group-Grain Striping at G=1. With current disk technology, having G=1 maximizes the number of streams. But as the disk access time drops, it is beneficial to increase G, so that G rounds worth of stream data are transferred in a single round.

Specifically, when using the mixed workload, it is anticipated that two years into the future, the number of streams that could be supported with Group-Grain policy at G=2 increases by 35% when compared to Fixed-Grain Striping. Five years into the future, the corresponding benefit of Group Grain Striping at G=3 remains at 29%. Thus, under reasonable technological improvements, there are significant performance improvements when using Group-Grain Striping instead of Fixed-Grain Striping.

Although preferred embodiments of the invention have been described herein, it will be understood by those skilled in the art that variations may be made thereto without departing from the scope of the invention. While this invention has focused on the striping problem for the common case of sequential playback of video, it will be understood by those skilled in the art that variations beyond the playback of video, such as, the download of a file or any other material or any stream data, may be made thereto without departing from the scope of the invention.

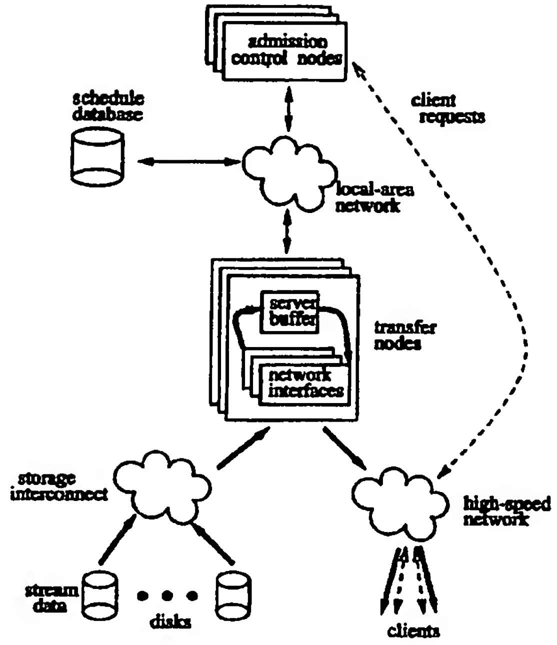


**Figure 1.** In the *Exedra* media server, stream data are retrieved from the disks and sent to the clients through the Transfer Nodes. Both the admission control and the data transfers make use of stream scheduling information maintained in the Schedule Database.

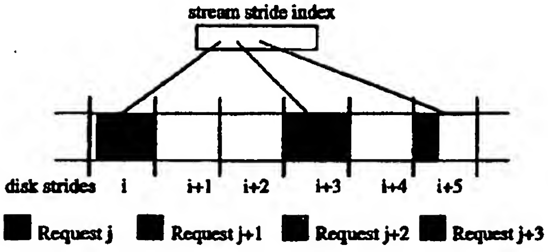


**Figure 2.** The stride-based allocation of disk space shown on one disk. A stream is stored in a sequence of generally non-consecutive fixed-size strides with a stride possibly containing data of more than one round. Sequential requests of one round are smaller than the stride size and thus require at most two partial stride accesses.

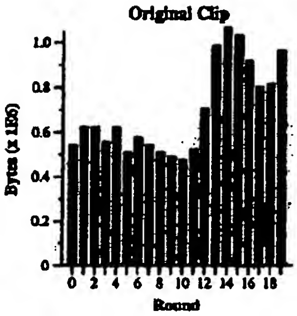
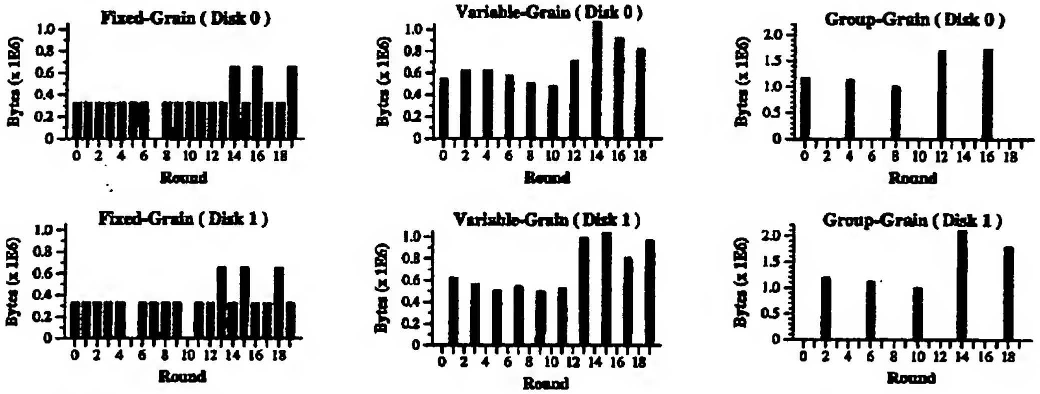


**Figure 3.** The data requirements of twenty consecutive rounds (one second each), in an MPEG-2 clip. Figure 4 shows how the clip is striped across two disks with alternative striping techniques.



(a) Fixed-Grain (Bf=327,680)

(b) Variable-Grain Striping

(c) Group-Grain (G=2)

**Figure 4.** Data accesses for the clip of Fig. 3 using alternative striping techniques over two disks. With Fixed-Grain Striping, the needed blocks of size $B_f$ are retrieved round-robin from the two disks every round. In Variable-Grain Striping, a different disk is accessed in each round, according to the byte requirements of the original clip. In Group-Grain Striping with G=2, stream data worth of two rounds are accessed from a different disk every two rounds.

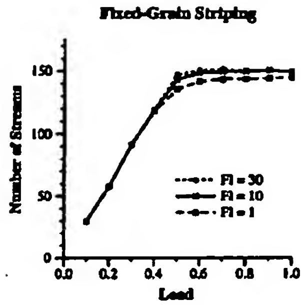


**Figure 5.** The sustained number of active streams with Fixed-Grain Striping flattens out at loads higher than 50% using a block size $B_f = 327,680$ with sixteen disks and the mixed workload. Changing the lookahead factor $F_l$ from 1 to 30 increases the number of streams by less than 5%.

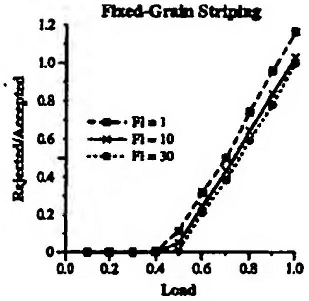


**Figure 6.** For Fixed-Grain Striping with $B_f = 327,680$, sixteen disks and the mixed workload, the total number of rejected streams over the total number of accepted during the measuring period. The ratio increases linearly as the load exceeds 50%, and changes by less than 20% as the lookahead factor $F_l$ is increased from 1 to 30.

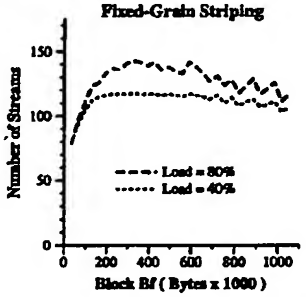


**Figure 7.** The number of active streams with Fixed-Grain Striping at different values of $B_f$. At both load values 40% and 80%, a maximum number of streams is achieved at $B_f = 327,680$. The experiments have been done over a range of $B_f$ between 32,768 and 1,048,576 at steps of 32,768. Sixteen disks have been used with the mixed workload.

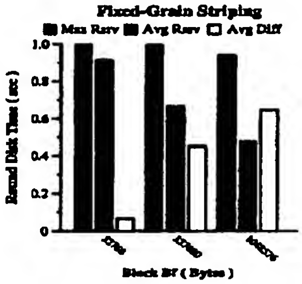


**Figure 8.** The maximum (Max Rsrv) and average (Avg Rsrv) disk access times reserved for a specific disk (Disk 0) in each round during the measuring period. Also, the maximum difference (Avg Diff) is shown between the reserved access times across all the disks in each round, averaged over the measuring period. Three different block sizes are tried with sixteen disks and the mixed workload at load $\rho = 80\%$.

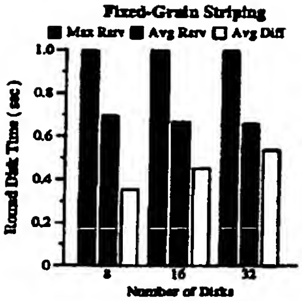


**Figure 9.** As the number of disks takes values 8, 16 and 32, the maximum and average reserved access time for Disk 0 (these values are respectively similar for the rest of the disks) remain almost the same. This is despite the corresponding increase in the maximum difference between reserved access times across the disks. The Fixed-Grain Striping policy is used on sixteen disks with mixed workload at load 80%.

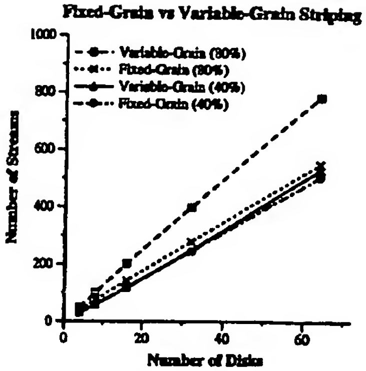


**Figure 10.** The sustained number of streams is measured using the Fixed-Grain Striping policy with a block size $B_f$ that maximizes the number of streams. At load $\rho = 80\%$, the streams increase almost linearly from 39.17 to 550.23 as the number of disks increases from 4 to 64, while for $\rho = 40\%$ the corresponding increase is from 30.31 to 504.79. With Variable-Grain Striping instead, the number of streams increases from 48.11 to 786.05 at $\rho = 80\%$, and from 29.86 to 530.09 at $\rho = 40\%$.

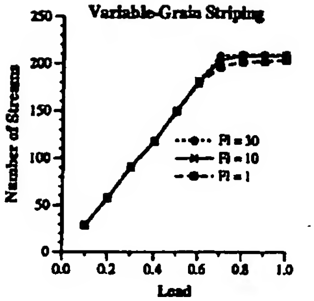


**Figure 11.** The sustained number of active streams with Variable-Grain Striping flattens out at loads higher than 70%. Changing the lookahead factor $F_l$ from 1 to 30 increases the number of streams less than 5%. Sixteen disks are used with the mixed workload.

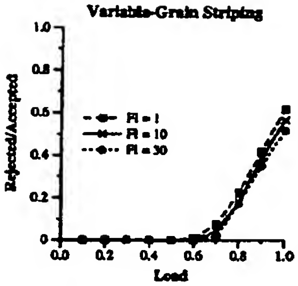


**Figure 12.** The total rejected streams over the total accepted during the measuring period. The ratio increases linearly as the load exceeds the 70% value, and changes by no more than 10%, as the lookahead factor $F_l$ varies between 1 and 30. Sixteen disks are used with the mixed workload.

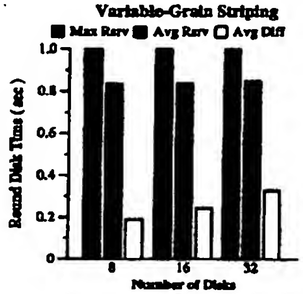


**Figure 13.** As the number of disks takes values 8, 16 and 32, the maximum and average reserved access time in Disk 0 (the values are respectively similar for the rest of the disks) remain almost the same. This is despite the corresponding increase in the maximum difference between reserved access times across the disks. The Variable-Grain Striping policy is used on sixteen disks with the mixed workload at load 80%.

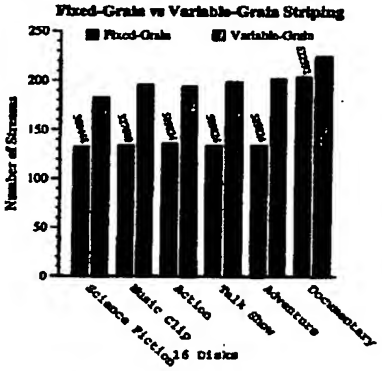


**Figure 14.** The advantage of Variable-Grain Striping over Fixed-Grain Striping varies among 38% in Science Fiction, 48% in Music Clip, 43% in Action, 49% in the Talk Show, 50% in the Adventure, and 11% in the Documentary. The block size shown for Fixed-Grain Striping was found to maximize the number of streams, over a range of block sizes between 32,768 and 1,048,576 with step of 32,768. The load was always set equal to 80%.

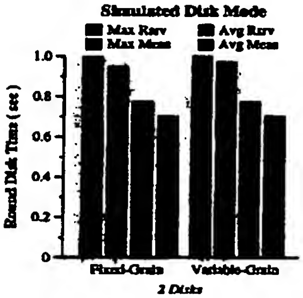


**Figure 15.** The reserved disk time statistics (Disk 0) are no more than 8% higher than the measured access time statistics using the detailed Seagate Cheetah ST-34501 model of Ganger et al.

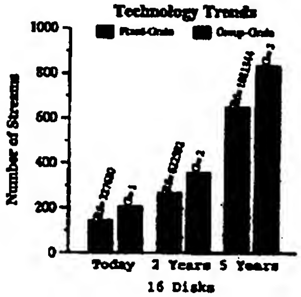


**Figure 16.** With reasonable technology projections, two years into the future Group-Grain Striping (generalised Variable-Grain Striping) maintains an advantage of 35% over Fixed-Grain Striping (from 41% today). The corresponding benefit in five years is no less than 29%. The shown values of $B_f$ and $G$ were found to maximize the throughput of the two policies respectively.